Reinforcement Learning Approaches to Instrumental Contingency Degradation in Rats

Alain Dutech, Etienne Coutureau, Alain Marchand

REINFORCEMENT LEARNING APPROACHES TO INSTRUMENTAL CONTINGENCY DEGRADATION IN RATS

Alain Dutech[1], Etienne Coutureau[2] et Alain R Marchand[2]
[1] LORIA/INRIA, Campus Scientifique, BP 239, 54506 Vandoeuvre les Nancy, France
[2] CNRS, CNIC-UMR5228, Bât. B2 av. Facultés, 33405 Talence, France
Alain.Dutech@loria.fr
E.Coutureau@cnic.u-bordeaux1.fr
A.Marchand@cnic.u-bordeaux1.fr

ABSTRACT

Goal directed action involves a representation of the consequences of an action. Rats with lesions of the medial prefrontal cortex do not adapt their instrumental response in a Skinner box when food delivery becomes unrelated to lever pressing. This indicates a role for the prefrontal region in adapting to contingency changes, a form of causal learning. We attempted to model this phenomenon in a reinforcement learning framework. Behavioural sequences of normal and lesioned rats were used to feed models based on the SARSA algorithm. One model (factorized-states) focused on temporal factors, representing continuous states as vectors of decaying event traces. The second model (event sequence) emphasized sequences, representing states as n-uplets of events. The values of state-action pairs were incorporated into a softmax policy to derive predicted action probabilities and adjust model parameters. Both models revealed a number of discrepancies between predicted and actual behaviour, emphasising changes in magazine visits rather that lever presses. The models also did not reproduce the differential adaptation of normal and prefrontal lesioned rats to contingency degradation. These data suggest that temporal difference learning models fail to capture causal relationships involved in the adaptation to contingency changes.

KEY WORDS
Rats, Instrumental, Contingency degradation, Simulation, Model-free learning

1. Introduction

Goal-directed behaviour requires a representation of the outcome of an action and an ability to adapt this action when its consequences change. In rodents as in humans, the prefrontal cortex contributes to both the acquisition and the flexibility of goal-directed instrumental behaviour [1]. Rats with lesions of the medial prefrontal cortex (mPFC) actually learn an instrumental task (lever pressing for a food reward) at a normal rate, but the response acquired is insensitive to contingency degradation, i.e. a weakening of the correlation between food delivery and lever pressing [2, 3]. In a design in which the outcome is equally probable in the presence or absence of an instrumental action, the mPFC is necessary to adapt to contingency degradation [4, 5]. The neural mechanisms of such a deficit in mPFC-lesioned rats are still poorly understood. Adaptation to contingency requires a learning process that integrates novel observations of unpredicted reward deliveries with a previously acquired action-reward association. As such, it should lend itself to modelling within the reinforcement learning framework [6]. Indeed, reinforcement learning processes occurring in the striatum have been proposed to underlie instrumental learning [7, 8]. Nevertheless, the role of the prefrontal cortex in this learning remains elusive [9, 10].

We have recently demonstrated the involvement of dopaminergic mechanisms within the prelimbic area of the mPFC in the adaptation to contingency changes [11]. Dopamine signals from ventral midbrain dopaminergic neurons are known to be modulated by uncertainty and delays in rewards delivery [12, 13]. Thus, new learning could be driven by the delivery of non-contingent rewards that occur in the absence of lever pressing and elicit a dopaminergic prediction error signal. A non-contingent reward implies that some time has elapsed or that events have occurred between the lever press and reward delivery, and we might expect mPFC-lesioned rats to exhibit deficits in maintaining a representation of their own actions in working memory against the passage of time or interfering events.

In the present study, we used temporal difference (TD) learning to test the hypothesis that prefrontal-lesioned rats have difficulties in parsing the flow of events so as to detect changing relationship between the rat's own actions and rewards. We examined this issue using a combined behavioural and simulation approach, with the following rationale: Behavioural data are first collected by training normal rats and rats with lesions of the mPFC in a standard operant task, followed by a contingency degradation phase. Then, a detailed analysis of behavioural sequences is conducted in order to identify differences in behaviour that might underlie deficits in adaptation to contingency changes. Finally, reinforcement-learning models are developed and trained using real event sequences, to determine whether different sets of parameters underlie the behavioural performance of normal and lesioned rats. Identifying such differences in model parameters would

provide valuable clues as to the operations performed in the region of interest, namely the mPFC.
Free operant learning raises special difficulties for reinforcement leaning because time is not divided into a series of discrete trials which would provide a natural support for the Markov processes on which the temporal difference algorithm is based [14]. We tested two models of free operant learning and contingency degradation that capture working memory constraints in different ways: 1) A factorized states model that emphasizes the temporal dynamics of parallel memory traces of past actions and events 2) An event sequence model that focuses on the span of working memory for successive events, irrespective of time. Both models were based on the SARSA algorithm and incorporated real sequences of actions and events to train the model and to adjust model parameters.

2. Methods and Results

- **Behavioural data:**

The behavioural experiments that served as basis for simulation involved a series of instrumental training sessions during which 12 rats bearing neurotoxic lesions of the mPFC and 16 control rats were trained when hungry to lever press for food pellets in a free operant task. This training phase consisted of two sessions of magazine training and 7 sessions of rewarded lever presses with progressively fewer rewards, down to an average of 1 reward/min (VI60 schedule). The rats were then switched to one of two possible new action-outcome contingencies [15]. In the negative contingency condition, the animals always obtain a food reward after a fixed time had elapsed when they do not press the lever (negative contingency). In the zero-contingency condition, reward delivery is yoked to that of an animal in the negative contingency condition, and thus independent of lever pressing. The results show that mPFC-lesioned rats are impaired in the zero-contingency condition, but not in the negative contingency condition, thereby demonstrating that mPFC-lesioned rats do not display an inflexible behaviour. Rather, they appear unable to detect weak contingency changes.

- **Modelling:**

The continuous nature of the task in free operant behaviour is a challenge for modelling, unless using the less tractable framework of semi-Markov Decision Processes [16]. At any instant, the rat may choose between various actions such as visiting the food magazine or pressing the lever. However, magazine visits will not lead to the same outcome depending on whether or not the rat has previously pressed the lever. This emphasizes the need to define states in this free operant situation. We chose to define states in two different ways: A factorized-states model that captures the temporal decay of working memory for past actions, independently of their order, and an event-sequence model that captures the limited capacity of working memory for sequences of consecutive events, without reference to their time of occurrence.

- **Temporal dynamics of traces:**

In the factorized-states model, a continuous state is defined at each time step as a vector of parallel event traces (actions and stimuli) that decay as time elapses. Reward prediction is computed as the inner product of the instantaneous state vector and a value vector corresponding to the action performed. The value of event-action pairs is updated according to SARSA-type learning with an eligibility trace for each action defined as a decaying trace of the state vector from the instant the action was performed. We examined whether normal and lesioned rats may be characterized by different decay times in their memory for actions. Four distinct actions were considered: pressing the lever (p), visiting the magazine (v), eating (e) or waiting (w). States were defined on the basis of the following events: lever presses (LP), visits to the empty magazine (ME), visits to the magazine with consumption of the reward (MR) and reward deliveries (RD) that correspond to the noise of the pellet dispenser.

- **Working-memory span:**

In the event-sequence model, states change only when a new event (stimulus or action) occurs, or when a predefined forgetting time has elapsed. States consist of n-uplets of events that correspond to the immediate past history of the animal. State-action pairs are updated according to the SARSA algorithm at every event occurrence. We examined whether normal and lesioned rats may be characterized by the number of successive events that can be held in working memory to constitute a state. We explicitly tested two cases: simple states consisting of the memory of a single event (the last event that has occurred), and complex states consisting of pair of events (the last two events experienced). In this model, we added a virtual event "forget" (FO) to model the notion that a rat could sometimes no longer take into account the last event if it happened too long ago in the past (characterized by a delay to forget parameter). The models allow an analysis of the real data in terms of the relative frequency of each action in the state defined by the model. To directly compare these frequencies with model predictions, the state-action values derived from the models at each instant were used to compute action probabilities using a softmax policy. The derived action probabilities were compared to observed action frequencies in order to adjust global model parameters for each group of rats.

- **Action probabilities in the various states:**

We first analyzed the rats' behaviour during the last session of instrumental training to determine how the frequency of the various actions depended on the state of the animal, as defined in each model.
In the factorized-states model, we computed an analog of conditional probability of action for each event trace as follows: action frequency weighted by trace intensity at the time of action occurrence, divided by the integral of the event trace over the whole session. The average values for lever presses and magazine visits during the

last session of training are represented in figure 1 for rats in the lesioned and control groups.

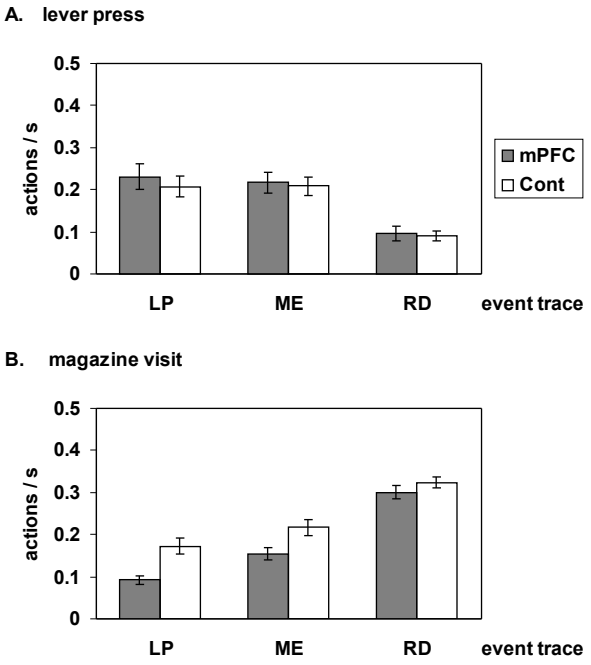

Figure 1. Action frequencies conditional on event traces. mPFC: lesioned group ; Cont: control group. LP: trace of lever press; ME: trace of visit to empty magazine; RD: trace of reward delivery. Time constant of event traces was 4 s.

This analysis shows that reward delivery (as signalled by the noise of the pellet dispenser) is an important event which induces a change of state and alters behaviour. Immediately after reward delivery, rats perform fewer lever presses and more magazine visits. In addition, the lesioned group appeared to perform fewer magazine visits than the control group in the absence of reward delivery.

In the sequence model, a similar analysis was conducted to determine whether rats were able to modulate their actions according to the preceding event sequence. Action probabilities were computed for each passage in each state, irrespective of the time spent in this state (Figure 2). The data indicate that both groups of rats modulate their behaviour not only after a reward delivery (LP&RD), but also after a waiting period (FO), as compared to a previous lever press or magazine visit. Specifically, rats appear to lever press more and perform fewer visits after waiting. However, probabilities of lever pressing and magazine visits are little affected by previous sequences of actions such as lever presses and magazine visits (FO&LP to ME&ME), suggesting that rats tend to remember only the last event or action. Again, the sole difference between normal and lesioned rats concerned the overall frequency of magazine visits. This analysis therefore does not reveal any clear difference between groups by the end of instrumental training.

The actual behavioural sequence of each rat was used to train each model across the successive sessions of training. As a result, values for each state-action pair were obtained and probabilities for each action in each state could be computed on a moment-by-moment basis using a softmax policy. Model parameters were adjusted by comparing the resulting predicted action probabilities to actual action frequencies over the final session of training.

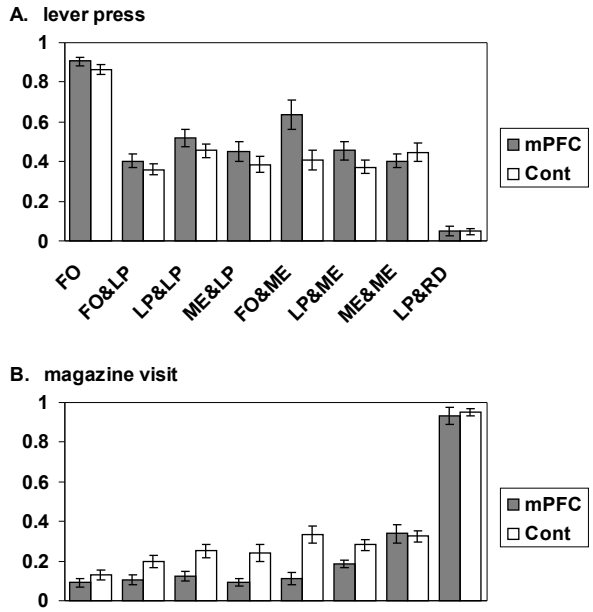

Figure 2. Action probabilities conditional on preceding event sequences. mPFC: lesioned group ; Cont: control group. Preceding events: FO: forgotten; LP: lever press; ME: visit to empty magazine; RD: reward delivery. Delay to forget: 5 s.

Figure 3 shows actual and simulated action frequencies in the factorized-states model with the same representation as in Figure 1.

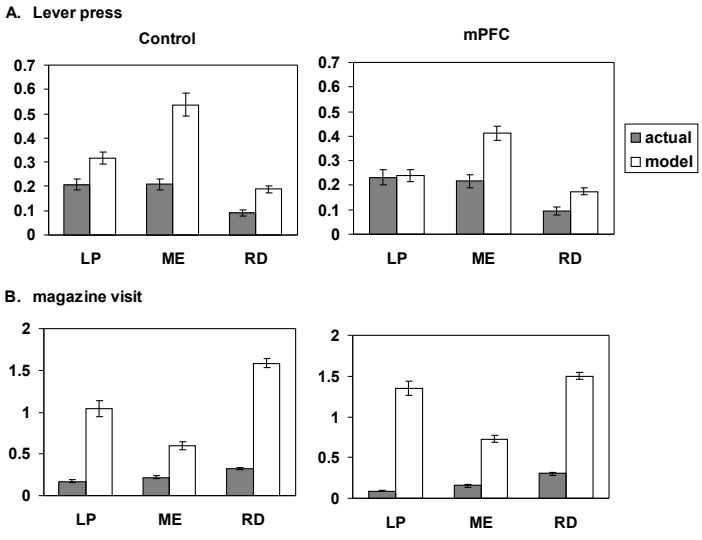

Figure 3. Actual and simulated action frequencies conditional on event traces. Representation as in Figure 1.

- **Performance of the models:**

Overall, the factorized-states model tends to overestimate action probabilities for all actions, possibly because of a difficulty in assigning a value to waiting. To some extent, the model captures the effects of reward delivery on the frequencies of lever pressing and magazine visits which was observed in behavioural data. However, the model consistently predicts that magazine visits should facilitate subsequent lever presses and reduce subsequent magazine visits, in contradiction to the observed data. In addition, no difference emerges between groups in the simulated data. A detailed analysis across training sessions indicates that the value of magazine visits following reward delivery tends to increase throughout training in the model, whereas the observed frequency of magazine visits reaches a plateau after 5 sessions. Moreover, the value of lever pressing is very sensitive to reward delivery rates, whereas actual rats do not press less when the rewards are made less frequent. These discrepancies between real and simulated data persist over a wide range of model parameters.

Predicted action probabilities in the sequence model are shown in Figure 4. As we showed above that behaviour was essentially modulated by one preceding event, these data are conditional on sequences of length 1, i.e. the last preceding event.

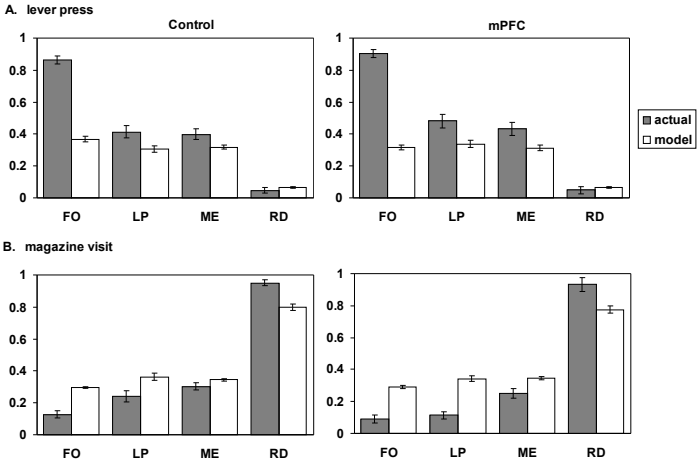

Figure 4. Actual and simulated action probabilities conditional on sequences of length 1 of preceding events. Same notation as in Figure 2. Model parameters were: learning rate: 0.4 ; temporal discount factor: 0.7 ; reward: 1 ; cost of lever press: -0.1 ; delay to forget: 5 s ; latency for action 'wait': 2.5 s ; inverse temperature for softmax: 2.

The sequence model appear to provide a much better fit of action probabilities than the factorized-states model, although it still tends to underestimate lever presses and to overestimate magazine visits. The model correctly captures the effects of reward delivery on the frequencies of lever pressing and magazine visits which was observed in behavioural data. However, it fails to reproduce the modulation of behaviour related to forgetting (FO). Importantly these effects were quite robust with respect to variations in model parameters. Again, both groups showed very similar results in the simulation as well as in the behavioural data.

During instrumental training, rats with lesions of the mPFC acquired the task and performed lever presses at a similar rate as control rats. However, both groups differed during the contingency degradation session where reward delivery was rendered independent of lever pressing (zero-contingency condition). Figure 5 shows the evolution of lever-press and magazine visit rates during contingency degradation in control and mPFC-lesioned rats. Only control rats reduce their lever presses over the duration of the session and this is accompanied by a moderate increase in magazine visits.

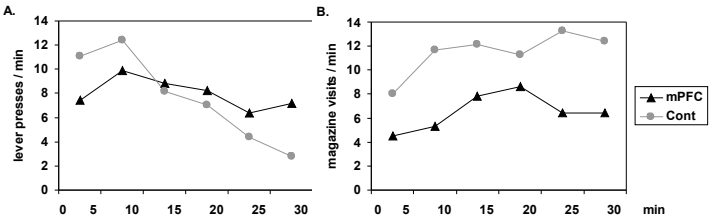

Figure 5. Rates of lever-presses and magazine visits over the duration of the contingency degradation session. Data were averaged over 6 successive blocks of 5 min. in control and mPFC-lesioned rats.

We applied both models to the contingency degradation session. With the factorized state model, the data were poorly fitted. The sequence model was applied with the same parameters as above. The resulting action probabilities are presented in figure 6.

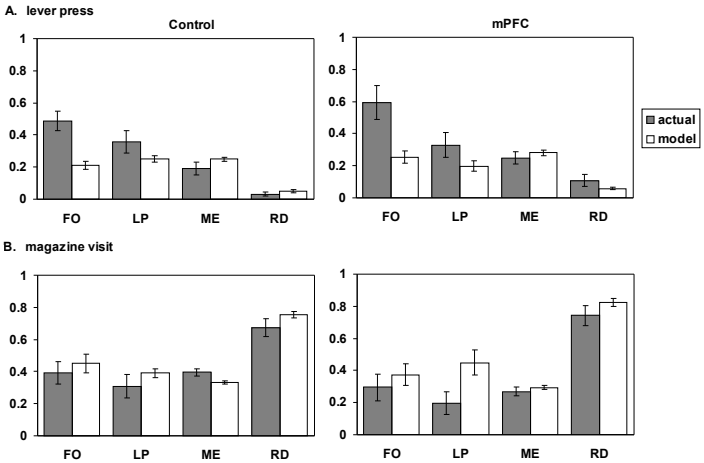

Figure 6. Actual and simulated action probabilities during contingency degradation, conditional on sequences. Same notation and model parameters as in Figure 4.

A comparison of Figures 4 and 6 shows that lever-press probabilities globally decreased and magazine visits increased in actual rats as a result of contingency degradation training. This effect was present even in mPFC-lesioned rats, although to a lesser degree. However, the large reduction in lever press frequency occurring in control rats does not clearly appear. The model however displays modest increases or decreases of action probabilities in the same direction as actual data. These data do not allow the model to differentiate mPFC-lesioned rats from normal rats. As previously, reward delivery appears a major determinant of the rat's actions. Clearly, the deficit of mPFC-lesioned rats in contingency detection is not obvious when considering action probabilities as in the sequence model. This is probably attributable to the fact that states defined in

the sequence model do not incorporate a regular definition of time.

3. Conclusion

- **Models in data analysis:**

This work aimed at contrasting, through two types of models, working memory and timing deficits in rats with prefrontal lesions. The results of this exploratory study demonstrate that simple models such as temporal difference learning are able to capture some aspects of performance in free operant behaviour. Importantly, one of the first steps in specifying such models is to define states for the system. This definition stage has major implications not only on the model's behaviour, but also in the way actual data can be considered and analyzed. In the present study, we compared two definitions of states, one emphasizing timing and parallel event traces, the other emphasizing event sequences. Both definitions, when applied to the analysis of actual behaviour of the rats, revealed a major role for an external event, i.e. the sound of reward delivery, in determining the rat's actions. Modelling sequences also allowed us to determine that rats do not modulate their behaviour according to sequences of preceding events or actions, but mainly take into account the most recent preceding event. Overall, predictions derived from the sequence model allowed matched actual action probabilities more closely than predictions derived from the factorized-states model.

The strategy proposed in this study can be summarized as follows: Defining model states provides a way to analyze behavioural sequences in order to identify differences in behaviour in various groups of animals. These states could be optimized by choosing relevant observable variables and adjusting parameters pertaining to hidden states, such as forgetting, with the criterion that actual behaviour should be as much as possible differentiated between states. When conducting simulations, these states allow a direct comparison of predicted and actual data. We chose here to use actual behavioural event sequences to feed the model, rather than letting the model follow its internal policy. The SARSA algorithm is particularly well suited to this procedure since state values only depend in this case on the next action performed, and there is no need to specify a theoretical policy. This strategy implies a break in the loop between acquired action values and the behaviour/policy of the model. We would like to emphasize that it represents a strong constraint for modelling since it reduces the effects of model parameters on its expected behaviour. Thus, obtaining a good fit under such constraints remains a challenge and would represent an important step towards biological plausibility and identification of brain operations. A final validation of this approach would be to close the loop by letting the model follow its internal policy.

- **Interpreting the function of mPFC:**

At the present stage of the study, the data do not allow a clear interpretation of mPFC function in instrumental behaviour. A deficit of mPFC-lesioned rats can be evidenced during a contingency degradation session where reward delivery is made independent of the previously causal action, namely lever press. The fact that this deficit is only apparent when considering response rates emphasizes the role of timing factors in this task. Specifically, the random time interval between lever-press and response delivery does not appear to lead to a perception of degraded contingency in animals without a functional mPFC, so that these animals do not reduce their response rates. These data are consistent with the putative involvement of prefrontal areas in cross- temporal associations [17], namely as a network that in the course of behaviour integrates information in a timely manner. mPFC-lesioned rats might thus be unable to perceive deviations from the normal temporal sequence. This perception may involve hidden states corresponding to the fact that previous events are forgotten or disregarded.

Further work will be required to test this type of hypothesis. So far, our models failed to capture the difference between these groups of animals. Specifically, the factorized-states model does not properly match behavioural data, so that it is difficult to evaluate whether mPFC-lesioned rats are deficient in their perception of delays to reinforcement. The sequence model revealed that mPFC-lesioned rats, but also normal rats, do not integrate sequences involving their own actions. It is therefore unlikely that their deficit should involve a shorter working-memory span than normal rats. Moreover, because of the crude way in which the sequence model encodes time, it cannot be used at present to differentially fit parameters to the two groups of animals.

- **Limits of the study:**

This study illustrates some of the difficulties in analyzing and modelling free operant behaviour with temporal difference learning models [18]. One of these difficulties is the continuous nature of the task which does not result in a natural definition of model states. The fact that the rats may perform any action at any point in time raises several problems: One of them is the difficulty in defining the absence of action, such as waiting. We chose here to consider that a single "wait" action occurred whenever the delay between successive actions exceeded a predetermined threshold (typically 2.5 s), but this remains somewhat artificial. A second difficulty resides in credit assignment in continuous time. Several actions or events may compete for predictive status and there are a potentially infinite number of time sequences preceding each reward. We selected to approach this problem in the factorized-states model by using eligibility traces and ignoring sequences, and in the sequence model by ignoring time altogether. Neither approach can be considered fully satisfactory. A third problem is the important role that emerged for reward delivery signals. During contingency degradation, rewards may occur independently of any action, so that this state transition is driven by an apparently unpredictable external event, whereas during training it is triggered by some of the lever presses. It is therefore a problem to define an action to be associated with the transition to cover both

the training phase and the contingency degradation phase.

However, it is the causal nature of contingency adaptation that is a major source of difficulty for reinforcement leaning. Contingency adaptation occurs when free rewards elicit a prediction error signal that should be used to update action values. However, the SARSA algorithm, like Q-learning and most TD algorithms, cannot update the value of an action when this action has not been recently performed. The contingency task assesses a rat's ability to take into account free reward deliveries so as to update the value of lever pressing (as lever presses are not needed any more and should be avoided). But because of the predictive nature of TD learning, no mechanism is provided to alter the value of lever presses in the absence of this action. Consequently, TD fails to capture the causal structure of the task and to provide mechanisms for fast adaptation of the instrumental response.

Whether simple models such as TD learning are sufficient to account for these properties remains to be determined. Indeed, TD learning as a model-free learning process may be usefully complemented by model-based systems that explicitly encode event consequences. The requirement of the mPFC to evaluate reward contingency agrees with computational models emphasizing the encoding of consequences and uncertainty processing in the PFC as a determinant of action choice [9]. Such models have received support from studies in monkeys and humans showing activity in prefrontal regions that track changes in contingency [19, 20].

There is still a need to develop and evaluate learning algorithms that are both biologically plausible and able to fully capture the causal structure of a task. The contingency degradation task appears to be an appropriate test for these algorithms. Because it is based on a free operant learning task and takes the rate of action production as a dependent measure, further work is needed to develop models that integrate a continuous flow of actions and events. This approach can be expected to shed new light on the principles and determinants of action-consequence relationships, as well as to contribute to new developments in the field of autonomous robots.

References

[1] B.W. Balleine & J.P. O'Doherty, Human and Rodent Homologies in Action Control: Corticostriatal Determinants of Goal-Directed and Habitual Action, *Neuropsychopharmacology*, *35*(148-69), 2010, 48-69.

[2] S. Killcross & E. Coutureau, Coordination of actions and habits in the medial prefrontal cortex of rats, *Cerebral Cortex*, *13*(4), 2003, 400-408.

[3] S.B. Ostlund & B.W. Balleine, Lesions of medial prefrontal cortex disrupt the acquisition but not the expression of goal-directed learning, *Journal of Neuroscience*, *25*(34), 2005, 7763-7770.

[4] B.W. Balleine & A. Dickinson, Goal-directed instrumental action: contingency and incentive learning and their cortical substrates, *Neuropharmacology*, *37*(4-5), 1998, 407-19.

[5] L.H. Corbit & B.W. Balleine, The role of prelimbic cortex in instrumental conditioning, *Behavioural Brain Research*, *146*(1-2), 2003, 145-157.

[6] R.S. Sutton & A.G. Barto, *Reinforcement Learning: An Introduction* (Cambridge, MA, London, UK): MIT Press), 1998).

[7] D. Joel, Y. Niv & E. Ruppin, Actor-critic models of the basal ganglia: new anatomical and computational perspectives, *Neural Netw.*, *15*(4-6), 2002, 535-547.

[8] H. Kim, J.H. Sul, N. Huh, D. Lee & M.W. Jung, Role of Striatum in Updating Values of Chosen Actions, *29*(47), 2009, 14701-14712.

[9] N.D. Daw, Y. Niv & P. Dayan, Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control, *Nature Neuroscience*, *8*(12), 2005, 1704-1711.

[10] M.J. Frank & E.D. Claus, Anatomy of a decision: striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal, *113*(2), 2006, 300-26.

[11] F. Naneix, A.R. Marchand, G. Di Scala, J.R. Pape & E. Coutureau, A role for medial prefrontal dopaminergic innervation in instrumental conditioning, *Journal of Neuroscience*, *29*(20), 2009, 6599-6606.

[12] C.D. Fiorillo, P.N. Tobler & W. Schultz, Discrete coding of reward probability and uncertainty by dopamine neurons, *Science*, *299*(5614), 2003, 1898-902.

[13] S. Kobayashi & W. Schultz, Influence of reward delays on responses of dopamine neurons, *Journal of Neuroscience*, *28*(31), 2008, 7837-46.

[14] N.D. Daw, A.C. Courville & D.S. Touretzky, Representation and timing in theories of the dopamine system, *Neural Computation*, *18*(7), 2006, 1637-1677.

[15] H.H. Yin, B.J. Knowlton & B.W. Balleine, Inactivation of dorsolateral striatum enhances sensitivity to changes in the action-outcome contingency in instrumental conditioning, *Behavioural Brain Research*, *166*(2), 2006, 189-96.

[16] S. Bradtke & M.O. Duff, Reinforcement learning methods for continuous-time Markov decision problems, *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 1995.

[17] J.M. Fuster, The prefrontal cortex-an update: time is of the essence, *Neuron*, *30*(2), 2001, 319-333.

[18] Y. Niv, N.D. Daw, D. Joel & P. Dayan, Tonic dopamine: opportunity costs and the control of response vigor, *Psychopharmacology (Berl)*, *191*(3), 2007, 507-520.

[19] K. Matsumoto & K. Tanaka, The role of the medial prefrontal cortex in achieving goals, *Current*

Opinion in Neurobiology, *14*(2), 2004, 178-185.

[20] S.C. Tanaka, B.W. Balleine & J.P. O'Doherty, Calculating consequences: brain systems that encode the causal effects of actions, *Journal of Neuroscience*, *28*(26), 2008, 6750-5.